A control theoretic model of adaptive behavior in dynamic environments

Harrison Ritz[1], Matthew R. Nassar[1], Michael J. Frank[1,2], Amitai Shenhav[1,2]

1. Cognitive, Linguistic & Psychological Science, Brown University, Providence, RI
2. Brown Institute for Brain Science, Brown University, Providence, RI

Correspondence: Harrison Ritz, Brown University, Providence, RI, 02912; email: harrison_ritz@brown.edu

A control theoretic model of adaptive behavior

Abstract

In order to behave adaptively in environments that are noisy and non-stationary, humans and other animals must monitor feedback from their environment and adjust their predictions and actions accordingly. An under-studied approach for modeling these adaptive processes comes from the engineering field of control theory, which provides general principles for regulating dynamical systems, often without requiring a generative model. The proportional-integral-derivative (PID) controller is one of the most popular models of industrial process control. The proportional term is analogous to the “delta rule” in psychology, adjusting estimates in proportion to each successive error in prediction. The integral and derivative terms augment this update to simultaneously improve accuracy and stability. Here, we tested whether the PID algorithm can describe how people sequentially adjust their predictions in response to new information. Across three experiments, we found that the PID controller was an effective model of participants' decisions in noisy, changing environments. In Experiment 1, we re-analyzed a change-point detection experiment, and showed that participants’ behavior incorporated elements of PID updating. Experiments 2-3 employed a gradual change-point task that was optimized to detect PID-like adjustments. In both experiments, the PID model offered better descriptions of behavioral adjustments than the classical delta-rule model. We further examined how participants weighted different PID terms in response to salient environmental events. Across experiments, we found that the weights on these control terms were modulated by reward, surprise, and outcome entropy. These experiments provide preliminary evidence that adaptive behavior in dynamic environments resembles PID control.

Keywords: learning, delta-rule, PID control, adaptive control, monitoring

A control theoretic model of adaptive behavior

To behave adaptively, we must adjust our behavior in response to the dynamics of our environment (Ashby, 1956; Pezzulo & Cisek, 2016). Achieving this goal requires us to collect feedback about the outcomes of our recent actions, and incorporating this feedback into decisions about how to adjust future actions. Within research on learning and decision-making, a popular approach for achieving this feedback-based control is referred to as the *delta-rule model*. This model adjusts expectations in proportion to the discrepancy between the outcome that was predicted and the outcome that occurred (i.e., prediction error), weighted by the *learning rate* ($\Delta expectation = learning\ rate \times error$; Widrow & Hoff, 1960; Rescorla & Wagner, 1972). While substantial evidence suggests that people implement some form of delta-rule control (Mirenowicz & Schultz, 1994), this algorithm has major limitations: first, it is sensitive to any noise that will cause persistent errors, either leading to oscillatory behavior (at a high learning rate) or a sluggish response (at a low learning rate; Aström & Murray, 2010). More seriously, this algorithm performs poorly in any environment that changes over time (Pearce & Hall, 1980; Aström & Murray, 2010).

More elaborate feedback control mechanisms have been developed for use in various industrial applications within a branch of engineering that studies the regulation of dynamical systems, referred to as Control Theory. Control Theory augment delta rule-like algorithms with additional control terms that greatly improve the efficacy of this algorithm. The most popular variant of these control theoretic models is referred to as the popular proportional-integral-derivative (PID) controller (Fig. 1). This model is simple, accurate, and robust, with response properties that have been well-characterized over the last century (Franklin, Powell, & Emami-Naeini, 1994; Aström & Murray, 2010). The PID controller takes the error from a reference signal as input, and it outputs a control signal consisting of a linear combination of control signals proportional to the error (P-Term), the integral of the error (I-Term), and the derivative of the error (D-Term; Fig. 1). These three terms minimize deviations from the reference based on errors in the present, past, and expected future, respectively.

Proportional control (cf. delta-rule control) directly minimizes deviation from the reference, and is often the primary driver of the control process. Integral control provides low-frequency compensation for residual steady-state errors, allowing the controller to reduce noise and track gradual changes in the environment. Derivative control provides high-frequency compensation that increases stability, dampening control adjustments when the controller is approaching the reference and increasing these adjustments when the reference or environment changes suddenly. Here we test

whether this popular model of industrial control can account for adjustments in human behavior within a dynamic environment.

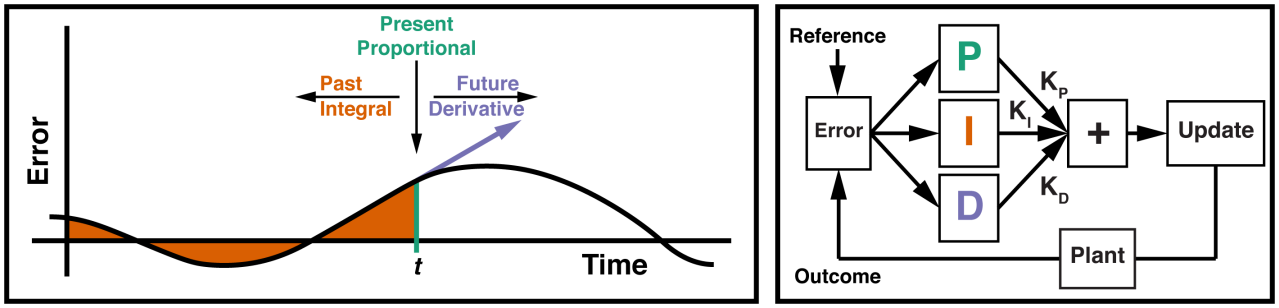

Figure 1. **Left:** The Proportional-Integral-Derivative (PID) Controller uses the current error (P-Term), the integral of the error (I-term), and the derivative of the error (D-Term), in order to provide compensation based on the present, past, and future, respectively. **Right:** The PID controller specifies a control signal based on the weighted sum of the PID terms, with each term weighted by their respective gain. Similar to a thermostat-controlled furnace, the plant implements the control signal, moving the measured processes closer to the reference (e.g., the desired temperature). Left panel adapted from (Aström & Murray, 2010), permission pending.

PID control has algorithmic properties that make it useful for most control systems. For instance, relative to algorithms that require an explicit representation of task dynamics, PID can provide an effective, and computationally cheaper, model-free alternative to adjusting cognitive or behavioral processes over time, particularly for natural environments that require particularly complex world models. Moreover, convergent evidence suggests that the PID algorithm may help account for the variety of feedback-related findings observed in humans and monkeys. Behavioral and neural correlates of feedback-controlled choice provide preliminary evidence that participants transform decision-relevant variables in a manner predicted by the PID algorithm. Consistent with proportional control, there is substantial evidence that participants adjust their behaviors based on recent errors or conflict (Rescorla & Wagner, 1972; Lau & Glimcher, 2005; Rabbitt, 1966; Gratton, Coles, & Donchin, 1992; Ullsperger, Danielmeier, & Jocham, 2014), with corresponding signals observed most prominently in the striatum and anterior cingulate cortex (ACC; Mirenowicz & Schultz, 1994; Garrison, Erdeniz, & Done, 2013; Niki & Watanabe, 1979; Ito, Stuphorn, Brown, & Schall, 2003; Matsumoto, Matsumoto, Abe, & Tanaka, 2007; Kennerley, Walton, Behrens, Buckley, & Rushworth, 2006; Seo & Lee, 2007; Smith et al., 2015). Consistent with integral control, experiments and simulations have suggested that people are sensitive to the extended history of errors or conflict (Laming, 1968; Logan & Zbrodoff, 1979; Botvinick, Braver, Barch, Carter, & Cohen, 2001; Aben, Verguts, & Van den

Bussche, 2017; Alexander & Brown, 2011, 2015; Bugg & Crump, 2012; Wittmann et al., 2016), and experiments have found neural signals in the prefrontal cortex and ACC that reflect this history (Carter et al., 2000; Blais & Bunge, 2010; Bernacchia, Seo, Lee, & Wang, 2011; Wittmann et al., 2016). Recent models of the ACC have emphasized the role that integrative, recurrent activity in this region play in executive control (Wang, 2008; Hunt & Hayden, 2017; Shahnazian & Holroyd, 2017). Finally, consistent with derivative control, prior work has found that participants track the environmental rate of change when making decisions, with associated neural correlates in the anterior prefrontal cortices and ACC (Behrens, Woolrich, Walton, & Rushworth, 2007; Jiang, Beck, Heller, & Egner, 2015; Bernacchia et al., 2011; Kovach et al., 2012; McGuire, Nassar, Gold, & Kable, 2014; Wittmann et al., 2016).

Despite the success of the PID model as a simple and effective algorithm for implementing control in other fields, and evidence of relevant neural signatures in circuits involved in adaptive control, PID has yet to be formally tested as a model of human adaptive behavior. In the current set of experiments, we directly tested whether a PID model can describe human performance in adaptive learning tasks. In Experiment 1, we re-analyzed a recent study that examined predictive inference in an environment with discrete change points. Behavior on this task confirmed key predictions of the PID model, but was limited in its ability to adjudicate between candidate models. Informed by our findings in Experiment 1, for Experiments 2-3 we developed a novel task that was optimized for PID control, using gradual rather than sudden change-points. We found that the PID model was a strong predictor of participants' choices in both experiments. Experiment 3 replicated the predictive power of our model, and further examined whether participants dynamically adjust their control terms based on rewards, surprise, and outcome entropy. Across these tasks, participant performance was consistent with key predictions of the PID model, demonstrating that this simple model provides a promising account of adaptive behavior.

Experiment 1

The PID model is designed to adapt the behavior of a system in response to changes in the environment. We therefore began by testing whether this model could explain behavioral adjustments in an existing change-point detection task, one that was designed to assess how humans can adapt their learning rate to uncertain and volatile outcomes (McGuire et al., 2014). In this experiment, participants predicted where a target stimulus would appear (horizontal location on the screen), and then received

feedback about where the true location of the outcome had been (See Fig. 2A). Outcome locations were normally distributed around a mean, and the mean of this distribution changed suddenly throughout the experiment (change points), according to a predetermined hazard rate. This task allows us to measure participants' choices and feedback in a continuous space with high precision, making it desirable for studying PID control. We can therefore use the PID model to predict trial-to-trial adjustments in participant behavior (predicted locations) based on their history of error feedback. In other respects, this task is not ideally suited for testing our model: the dramatic changes in target distributions may 'reset' adaptive control processes (Bouret & Sara, 2005; Tervo et al., 2014), and so this experiment serves as a preliminary test of our hypothesized control dynamics. We will address these concerns in Experiments 2-3.

Methods

Participants and Procedure

Experiment 1 consisted of a re-analysis of a change-point task used by McGuire and colleagues (2014). Briefly, thirty-two participants (17 female, mean(SD) age = 22.4(3.0)) selected a location on a line using a joystick, and then were shown the correct location for that trial. On a random subset of trials, participants were rewarded if their accuracy was within a criterion, which de-correlated trial-wise errors and rewards. Target locations were drawn from a Gaussian distribution with a variable mean and variance. The mean was stable for three trials, and then on a weighted coin flip (hazard rate: 0.125) was uniformly redrawn from the line coordinates; the variance alternated between high and low levels across blocks. Participants performed 160 training trials followed by 4 blocks of 120 trials during fMRI. Two participants were excluded for having an inconsistent number of trials per block, leaving 30 participants for the final analysis. See McGuire et al. (2014) for additional details.

Lagged Regression Analysis

A critical prediction of the PID model is that a participant's updates (i.e., the change in their location guesses) should depend on their history of feedback. While a delta-rule model predicts that only the error on the current trial would directly influence updates, a PID controller integrates errors over a longer history, enabling the controller to correct for a consistent bias in errors. In order to measure the independent contribution of each trial's feedback in the recent past, we regressed the update in location estimates between two trials (current vs. previous) on the errors from the

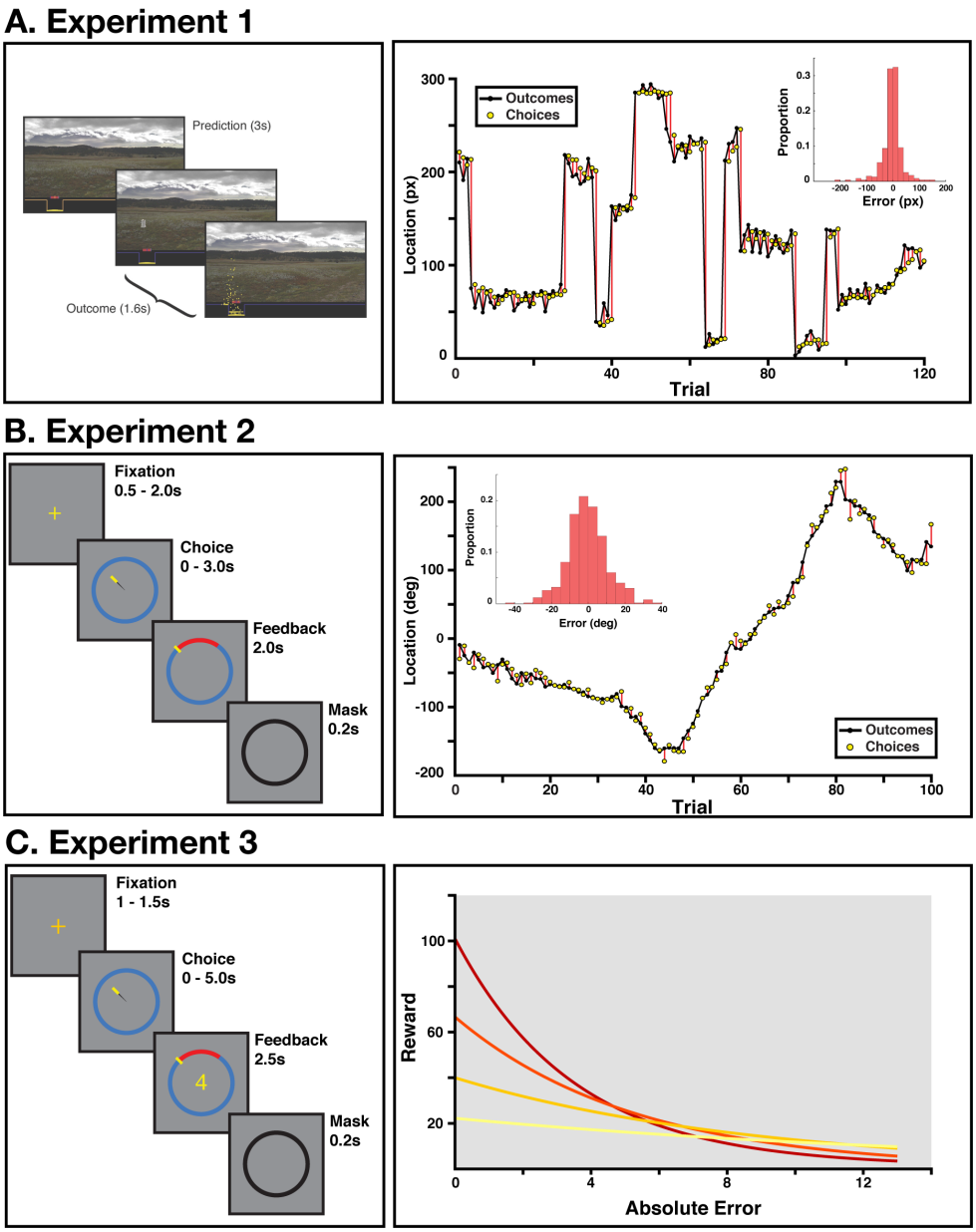

Figure 2. Experimental tasks. A) Left: On each trial of Experiment 1, participants selected a horizontal location with a joystick, and then were shown the correct location. Figure adapted from McGuire et al., 2014, permission pending **Right:** A representative block of trials from an example participant. The mean correct location was stable for a variable number of trials, and then was uniformly resampled. Inset histogram: The distributions of errors for this participant across their session. **B) Left:** Participants in Experiment 2 selected a location on the circle with their mouse, and then were shown the correct location**. Right:** A representative block of trials, demonstrating that the mean correct location changed gradually over time. As seen in the histogram for an example participant (inset), the gradual changes in location for this task resulted in error distributions that were less leptokurtic than in Experiment 1 (compare Panel A inset). **C) Left:** Experiment 3 was identical to Experiment 2 but participants were rewarded based on their accuracy, according to one of four reward-error functions. They were informed of the current reward mode during Fixation, and during Feedback they received the reward corresponding to their accuracy on that trial (conditional on their reward mode). **Right:** Error-reward slopes for the four potential reward modes.

current and 10 leading trials ($u_t \sim 1 + e_t + e_{t-1} + \cdots + e_{t-10}$; Wilkinson & Rogers (1973) notation) We assessed the influence of pervious trials' feedback by testing whether the sum of previous trials' betas was significantly different from zero, using a nonparametric sign-randomization test at the group level (comparing the observed results to a null distribution that we generated by randomly assigning positive or negative signs to each participant's summed betas). Throughout the paper, all randomization tests used 10^5 simulations, and all statistical tests were two-tailed with α = .05.

PID Model

The PID algorithm controls a process in order to maintain a desired reference signal (Fig. 1). It takes as input the signed error relative to this reference (e_t = reference - output), and produces a control signal (u_t) that specifies the adjustment for the next time point (here, the next trial). The control signal is defined by a linear combination of three terms: the P-term (reflecting the error), the I-term (reflecting the leaky integral of the error), and the D-term (reflecting the derivative of the error). Each of these terms is weighted by its own gain parameter (K_P, K_I, K_D). For trial t, the control signal (u_t) is therefore determined by transforming the error (e_t) as follows:

$$u_t = K_P e_t + K_I \sum_{n=1}^{t} \lambda^{t-n} e_n + K_D (e_t - e_{t-1})$$

where λ represents a memory persistence parameter, with larger values leading to longer retention. On the first trial of a block, the I-Term is e_t, and the D-Term is 0, producing a control action similar to proportional control. In the following tasks, u_t was defined as the difference in the choice location between trial $t+1$ and trial t (hereafter, the 'update'), and e_t was the difference between the correct location and the chosen location.

PID Model Fit

We used each participant's timecourse of errors within each block to generate hypothesized P, I, and D values based on the raw errors, the integral of the errors, and the first derivative of the error, respectively. Our regression model consisted of an intercept and the three PID terms ($u \sim 1 + P + I + D$), and we fit this model with iteratively reweighted robust regression (using MATLAB's fitlm function; bisquare weighting factor), in order to minimize overfitting to the outliers that can occur when participants make scalar responses. Fit statistics were generated based on the non-

reweighted residuals from the robust model, in order to avoid undue bias in favor of complex models.

Because the λ parameter (memory persistence) interacted with the identifiability of our PID terms when estimated jointly (e.g., when λ = 0, P and I are identical), we chose to fit this single term at the group rather than individual level. We fit λ with a grid search (range: [0.5-1], increments of 0.001), using median R^2 across participants as our measure of fit (in order to normalize individual differences in update variability). Regression models were estimated at the individual level, and regression weights were tested for deviance from zero at the group level with a sign-randomization test (see above).

We compared the P (i.e., delta-rule), PI, PD, and PID models, as these are the most common instantiations of the PID algorithm. To compare model performance, we used each participant's Akaike's Information Criterion (AIC; Akaike, 1983), an index of goodness-of-fit that is penalized for model complexity[1]. We compared the AIC at the group level using Bayesian model selection (Rigoux, Stephan, Friston, & Daunizeau, 2014), quantifying the Bayesian omnibus risk (likelihood that all of models are equally good) and each model's protected exceedance probability (PXP; the likelihood that the proportion of participants generated using a given model exceeds the proportion of participants generated using other models, controlling for the chance rate).

PID Controller Simulations

To determine whether participants were behaving similarly to the best performing PID controller for this environment, we simulated PID controller performance for the outcome locations that each participant encountered during the experiment. We used a restricted maximum likelihood estimation procedure (MATLAB's fmincon) to determine the values of the PID gains, λ, and choice bias (i.e., intercept) that perform best given the outcomes of each participant's task. We then compared these best-performing PID gains to the gains estimated from participants' behavior.

We also tested whether our fitted model would produce the same pattern of behavior that we measured with our lagged regression. We simulated an ideal observer that used each participant's estimated PID parameters and outcome history in order to generate a sequence of updates, and then fit our lagged regression to this simulated behavior, separately for the P, PI, and PID candidate controllers. This analysis allows

[1] We chose AIC over the more conservative Bayesian Information Criteria (BIC) because model recovery found that BIC could not accurately recover the full PID model when it was the ground truth

us to qualitatively determine the extent to which the PID model can act as a generative model of participants' decision-making behavior (Gelman, Meng, & Stern, 1996; Nassar & Frank, 2016). If participants are using PID control, then simulated updates from a PID controller should similarly weight the feedback received over previous trials.

Results

Model-Agnostic Analysis

To identify the degree to which behavioral adjustment were influenced by recent feedback, we regressed participants current and previous errors on their update. We found that the current trial was the strongest predictor of the current update, but that updates were also influenced by feedback from leading trials (Fig. 3A). The sum of leading trials' betas was reliably less than zero (Mean(SD) summed beta: -0.040(0.064), p = .00017). This suggests that while immediate feedback was the most influential factor for participants' updates, they also incorporated an extended history of error feedback into their adjustments. While the current trial had a positive influence on updates, these previous trials instead had a negative influence on the current update, potentially compensating for the extreme errors that participants made at a change-points (see Fig. 2A).

To verify that our model can generate performance that captures the behavior observed in this task, we simulated behavior on this task using parameters estimated for our PID model and reduced versions thereof (P, PI). We then performed the same lagged regression on these simulated data that we used on real data (Fig. 3A). As expected, we found that the simulated PI and PID models captured the influence of leading errors, unlike the P-only model (which predicts that there should be no influence of leading errors).

PID Model Fit

We first simulated the best-performing PID agent for each participant's outcome history, finding that the simulated PID gains were all reliably different from zero (Mean(SD) PID Gain: K_P = 1.0(0.18), K_I = -0.16(0.12), K_D = -0.070(0.077), λ= 0.85(0.071); randomization test; all ps ≤ 10^{-5}; see Fig. 3B). Consistent with the lagged-regression analysis, we found that the simulated integral and derivative gains were negative.

Experiment 1

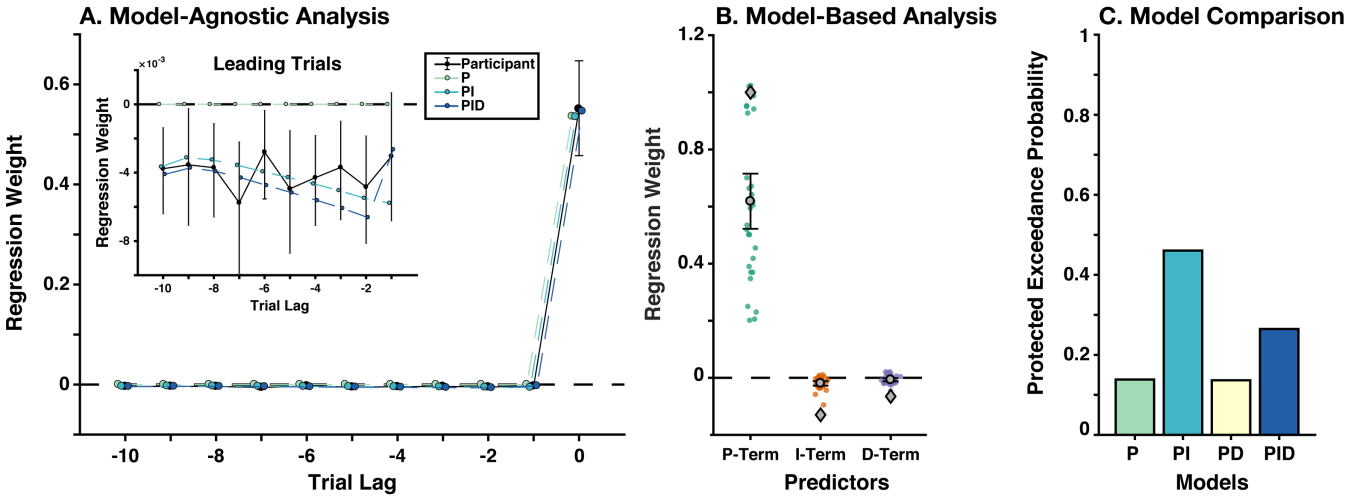

Figure 3. Experiment 1 Results. ***A)*** *For our model-agnostic analyses, we regressed the errors that participants made on the current and ten leading trials on their current update (black: participants' regression weights). Next, we used the PID parameters estimated in our regression analysis to generate behavior from P, PI, and PID controllers, and fit our lagged regression to this simulated behavior (colored lines). Inset: the regression weights from only the leading errors, controlling for the effect of the current error.* ***B)*** *For our model-based analyses, we regressed the trial-wise P-, I-, and D-Terms on participants' updates, and found that all three terms were significantly different from zero. Colored circles indicate individual participants' regression weights. Gray diamonds indicate the mean gains from ideal PID controllers fit to each participant's outcome history.* ***C)*** *We used Bayesian model selection to adjudicate between our candidate models, and found that the PI model best explained the data, albeit with moderate support (see text). Error bars throughout indicate mean and between-participant bootstrapped 95% confidence intervals.*

Fitting our PID model to participants' updates, we found that the best-fit models accounted for a substantial amount of this variance (median R^2 = 0.92), with parameters for all terms being significantly different from zero (Mean(SD) standardized betas: β_P = 0.62(0.27), $p \leq 10^{-5}$; β_I = -0.019(0.022), $p \leq 10^{-5}$; β_D= -0.0068(0.016), p = 0.022; see Fig. 3B). The group-level λ (memory persistence) was also relatively high (0.9430), suggesting that participants retained a great deal of information regarding past feedback. Participants' estimated gains qualitatively resembled the gains produced by the simulated PID controller, sharing the same sign and approximate magnitude (compare gray diamonds and circles in Fig. 3B).

Of these three models, we found that the PI model had the highest protected exceedance probability (PXP_P = 0.14, PXP_{PI} = 0.46, PXP_{PD} = 0.14, PXP_{PID} = 0.26; Fig. 3C), but that there is altogether insufficient evidence to support one model over another (Bayesian omnibus risk= 0.55, providing moderate evidence for the 'null' explanation that all models are equally likely). These data therefore do not allow us to

rule out the possibility that a simple delta-rule (P-only) model parsimoniously accounts for participant behavior.

Discussion

We found preliminary evidence that participants performing a change-point detection task are influenced by their history of error feedback, consistent with the predictions of a PID controller. Participants' updates could also be predicted from the integral and derivative of their errors. Despite these promising indications of PID control, we were unable to confidently differentiate between candidate models.

While this experiment offers mixed evidence in favor of the PID algorithm, this may be because this task was designed for change-point models, and has sudden, dramatic shifts in the outcome distribution. These change-points introduces extreme errors, which participants might treat categorically differently from normal prediction errors, evoking a 'reset' in their decision process or the representation of a different context (Bouret & Sara, 2005; Courville, Daw, & Touretzky, 2006; Nassar, Wilson, Heasly, & Gold, 2010; McGuire et al., 2014; O'Reilly et al., 2013). With this concern in mind, we developed a novel adaptive learning task in which outcomes changed smoothly over time, encouraging participants to treat outcomes as arising from a single, changing context.

Experiment 2

While Experiment 1 provided promising evidence that participants use their history of feedback in a way that resembled PID control, it did not provide definitive evidence as to whether this is the best explanation for the data. However, this may have been caused by participants strategically resetting their predictions at extreme change-points, making it more difficult to measure history-dependent predictions. To address this, we develop a *gradual* change-point task in which participants tracked outcomes distributions whose mean gradually changed from one location to another, and whose variance would change randomly throughout the block. In order to make these location transitions seem more continuous, they were represented along a circle rather than a straight line, thus also avoiding edge effects that can occur at either end of a screen. This design allowed us to precisely measure participants' predictions, errors, and adjustments within an environment whose dynamics are more fluid and predictable than Experiment 1. This task was explicitly designed to emulate an environment for which a PID controller is well-suited, and specifically to maximize our power to detect differences between PID control and proportional (delta-rule) control.

Methods

Participants & Procedure

Twenty-nine Brown University undergraduate students (25 female; mean(SD) age = 18.6(0.825)) performed a supervised learning task in which they predicted an outcome location on a circular display (see Fig. 2B).

Participants completed 5 blocks of 100 trials in which they used a mouse cursor to guess a location on the circumference of the circle. They were then shown the correct location, with an arc indicating the magnitude and direction of their error. Participants completed 50 training trials before the main experiment. Participants had up to 3 seconds to make their guess, or else their final cursor angle would be chosen as the guess for that trial, and feedback was presented for 2 seconds. Our final analysis excluded any trials where participants did not move their cursor to the edge of the circle, as well as a subset of trials following aberrant feedback due to a technical issue (1.8% of total trials).

The target location for each trial was drawn from a Gaussian distribution over arc degrees, with a mean and standard deviation that systematically changed over time. On a weighted coin flip (hazard rate: 0.80), the distribution's mean shifted based on a random draw from $U(-180, 180)$ degrees. After the new mean was drawn, the mean would transition from the old mean to the new mean over $U(8, 20)$ trials, with the means during transition trials linearly interpolated between the old and new means. The standard deviation varied independently of the mean, and was redrawn from $U(1, 8)$ degrees on a weighed coin flip (hazard rate: 0.40). These task parameters were selected through simulation in order to maximally differentiate the performance of PID and delta-rule models. Unless otherwise indicated, methods of analysis and model selection for this study are identical to Experiment 1.

Results

Model-Agnostic Analysis

Regressing the current and ten leading errors onto the current update (see Experiment 1 Methods), we again found that the sum of leading errors was significantly different from zero (Mean(SD) summed betas: 0.29(0.34), $p \leq 10^{-5}$; Fig. 4A). This replicates the observation in Experiment 1 that participants incorporate the extended history of errors into their prediction process.

Fitting our lagged model to behavior generated from our models produced a similar pattern of predictions as in Experiment 1: the P-only model categorically failed to capture the influence of leading errors. The PI and PID models were similar in their

ability to recreate participants' use of previous errors, however the PID model seemed to better capture participants' weighting of recent leading errors (i.e., over the previous three trials). We examined whether this trend was reliable across participants by fitting the linear and quadratic trend over trials to each participants' leading betas. We found a significant quadratic trend (Mean(SD) quadratic trend beta: -0.26(0.28), $p < 10^{-4}$) but not a linear one (Mean(SD) linear trend beta: 0.020(0.56), $p = .15$). While this finding is broadly compatible with derivative control's non-linear weighting of previous errors, the observed trend extended further backwards in time than the previous-trial effects predicted by a simple derivative.

Experiment 2

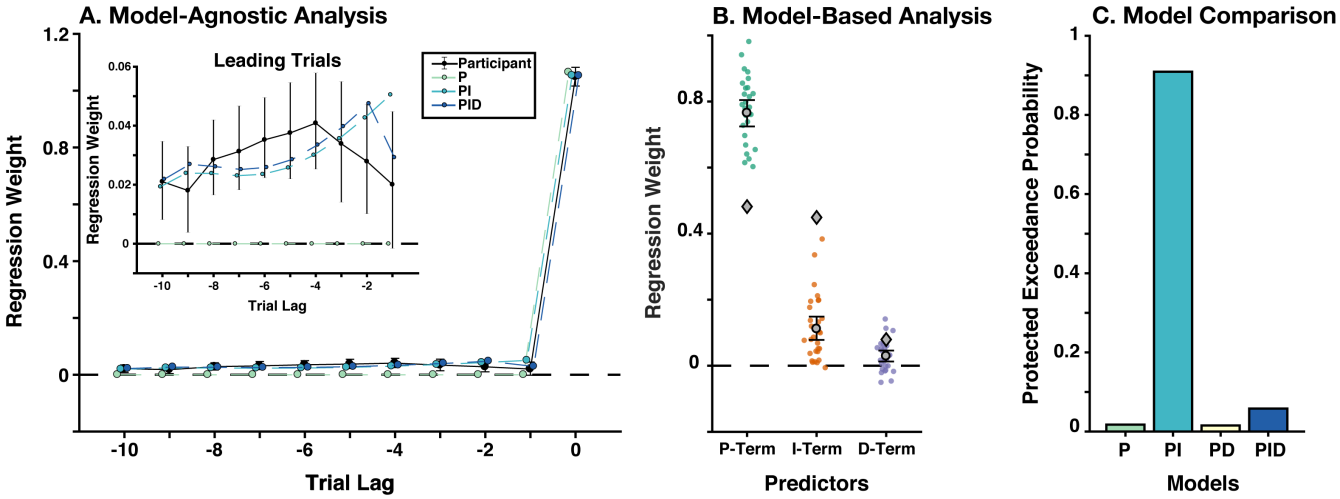

Figure 4. Experiment 2 Results. A) *Participants adjusted their choices based on previous errors, unlike the predictions of a proportional controller (i.e., delta-rule model).* ***B)*** *P-, I-, and D-Terms significantly predicted participants' updates.* ***C)*** *The PI controller best explained participants' behavior. See Figure 3 for detailed graph legends. Error bars indicate mean and between-participant bootstrapped 95% confidence intervals.*

PID Model Fit

The best performing PID gains were significantly different from zero (Mean(SD) gain: K_P = 0.45(0.13), K_I = 0.46(0.085), K_D = 0.089(0.064), λ = .88(0.034); all *p*s ≤ 10^{-5}; Fig. 4B). Our PID model accounted for most of the variance in participants' updates (median R^2 = 0.84), and the parameters for the P, I, and D-Terms were all significantly different from zero (Mean(SD) standardized beta: β_P = 0.77(0.11), $p \leq 10^{-5}$; β_I = 0.11(0.092), $p \leq 10^{-5}$; β_D = 0.029(0.050), p = 0.0017; sign-randomization test; Fig. 4B). The group-level λ was .860. Participants' parameters were similar to the ideal PID controller, although

they had a greater reliance on proportional control and a lesser reliance on integral control than was optimal.

Our model comparison found that there was a marginal difference between our models (Bayesian omnibus risk = 0.061), and Bayesian model selection favored the PI model (PXP_P = .018, PXP_{PI} = .91, PXP_{PD} = .015, PXP_{PID} = .058; Fig. 4C). While there was only a marginal omnibus difference between models, this was mostly due to the similarity in likelihood between the PI and PID model (excluding PID: Bayesian omnibus risk < .001, PXP_{PI} > .99). Therefore, our model selection provides strong evidence that an elaborated PID variant explains behavior better than strict proportional control.

Discussion

Using a novel variant of a change-point task, we provide strong evidence that the PID control model can usefully describe participants' predictions. Our model-free analysis showed that participants incorporated previous errors in their adjustments, in a way that is not predicted by a proportional control model, but is well approximated by our PID model. We also found that all of the PID terms were significant predictors of participants' updates, and that they were qualitatively similar to the gains of an ideal PID controller. Building on Experiment 1, this experiment provided strong evidence that PI was the best-fitting model. These data further support a role for control processes that extend beyond immediate errors.

These first two experiment have provided promising evidence that the PID controller predicts adaptive behaviors better than the classical delta-rule model. A striking feature of these experiments is that participants had very different estimated control gains across the two experiments, consistent with the differential gains of the best-performing PID agents. These differences suggest that participants may set their control gains in a context-specific manner, although at an unknown timescale. Popular delta-rule models have suggested that participants may in fact rapidly adjust their control gain in response to changes to the local context (Pearce & Hall, 1980). This prompted us to develop a third experiment, in order to replicate our results from Experiment 2, and test whether participants can adaptively PID gains to their local contexts.

Experiment 3

In Experiment 3, we sought to replicate the findings from Experiment 2, while at the same time manipulating the incentives for performing accurately on the task. We

additionally sought to examine three factors that might influence the weights that individuals might place on each of the PID terms over the course of an experiment.

First, we examined the influence of surprise (absolute error) on these control weights, given classic findings that such surprise signals modulate learning, indicating the degree to which the environment has been learned (Pearce & Hall, 1980; see also Hayden, Heilbronner, Pearson, & Platt, 2011; Nassar et al., 2010; McGuire et al., 2014; O'Reilly et al., 2013). Second, given evidence that learning can be influenced by uncertainty over recent feedback (Yu & Dayan, 2005; Courville et al., 2006; Nassar et al., 2010), or related estimates of volatility (Behrens et al., 2007), we examined how PID gains were influenced by an index of the outcome entropy over the past several trials. This measure of uncertainty indexes both expected uncertainty (the variance in the generative distribution) and unexpected uncertainty (changes in the mean of the generative distribution, i.e., drift), the latter of which is more dominant in our tasks. Finally, we examined the influence of reward on PID gains, given previous evidence that these can impact learning in a dissociable fashion from surprise or uncertainty alone (McGuire et al., 2014), and more generally that rewards may compensate for the costs of effortful control policies (Hayden, Pearson, & Platt, 2009; Padmala & Pessoa, 2011; Manohar et al., 2015; Kool, Gershman, & Cushman, 2017), including learning in particular (Hayden et al., 2009; Shenhav, Botvinick, & Cohen, 2013). Importantly, Experiments 1 and 3 were unique in the literature in that they were designed to de-confound reward from errors, providing us the ability to measure their influences on PID gains separately from one another and from our measure of uncertainty.

Methods

Participants and Procedure

Forty-seven Brown University subject pool participants (32 female; mean(SD) age = 21.30(4.07)) performed a rewarded supervised learning task (without monetary compensation). Apart from the reward manipulation, the structure of this task was similar to Experiment 2. On each trial, the reward magnitude depended on the accuracy of the participant's guess (i.e., the absolute error between guess and outcome location; Fig. 2C). These rewards decreased exponentially with increasing error magnitude. In order to de-correlate rewards and errors, and to vary overall motivation to perform the task, we adjusted the steepness (mean) of this exponential (gamma) function over trials, resampling one of four possible means (1, 1.5, 2.5, 4.5) at random time points, chosen with a flat hazard rate of .20 across all trials (Fig. 2C, right). These different levels of steepness defined four 'reward modes,' which participants

were instructed about at the beginning of the experiment. The reward mode for a given trial was indicated by the color of the fixation cross (one of four colors from equally spaced locations on a heat colormap). The input (errors) to these reward functions were divided by 3.5 in order to approximately match the reward that these functions returned to the mean performance level of Experiment 2.

Participants completed 50 training trials, followed by 6 blocks of 75 trials. On each trial, participants had up to 5 seconds to make their guess, feedback was presented for 2.5 seconds, and then the reward mode for the next trial was displayed during an ITI that was drawn from $U(1, 1.5)$. At the end of each block, participants were shown the mean reward earned during that block. Our final analysis excluded any trials where participants did not move their cursor to the edge of the circle (.07% of total trials). The lagged and trial-wise regression analyses were performed as described in Experiments 1 and 2.

Gain Modulation Analysis

In order to examine the influence of reward (Experiments 1 and 3; n = 77), errors (Experiments 1-3; n = 106), and outcome entropy (Experiments 1-3; n = 106) on the gains of the PID terms, we re-ran our PID regression analysis, including interaction terms for gain modulation. In Experiment 1, the reward modulator consisted of binary reward feedback that was given on a random subset of trials, conditional on participants' error being within a pre-specified threshold. This feedback was orthogonal to absolute error on the task. In Experiment 3, the reward modulator was the number of points that participants received on each trial. In all three experiments, the error modulator was the absolute prediction error. In all three experiments, outcome entropy was defined as the natural logarithm of the outcome sample standard deviation over the current and 10 previous trials within each block (with a truncated window for the first 10 trials in each block). In Experiment 3, participants received both error and reward feedback on every trial, and the absolute error was moderately correlated with the reward amount (median $r = -.68$). We ran Belsley collinearity diagnostics (Belsley, Kuh, & Welsch, 1980), and found that the collinearity between absolute error and reward was below standard tolerances.

A robust regression (bisquare weighted) was run for every participant in every experiment, with the reward modulator excluded for Experiment 2. The regression model included all main effects, as well as the interactions between the PID terms and gain modulators ($u \sim 1 + (P + I + D) * reward + (P + I + D) * absolute\ error + (P + I + D) * outcome\ entropy$). We mean-centered betas within their respective

experiment and then re-centered the betas on their grand mean, in order to remove between-experiment variance (Cousineau, 2005).

Results

Model-Agnostic Analysis

Regressing the current and ten leading errors onto the current update (see Experiment 1 Methods), we replicated the observation that participants were influenced by past errors (Mean(SD) summed betas: 0.25(0.21), $p \leq 10^{-5}$; see Fig. 5A). Our model-generated behavior again showed that the delta-rule model categorically fails to capture the influence of leading errors. Unlike Experiment 2, here we found that the weighting of previous errors was best fit as a linear decay from the current trials, resembling PI control (Mean(SD) trend beta: linear = 0.24(0.35), $p < 10^{-4}$; quadratic = -0.03(0.39), p = .40; sign-randomization test). This discrepancy from Experiment 2 may be due to the fact that rewards in Experiment 3 were highly dependent on accuracy. This may have biased participants more toward integral control (which favors accuracy) and away from derivative control (which favors stability; Aström & Murray, 2010).

Experiment 3

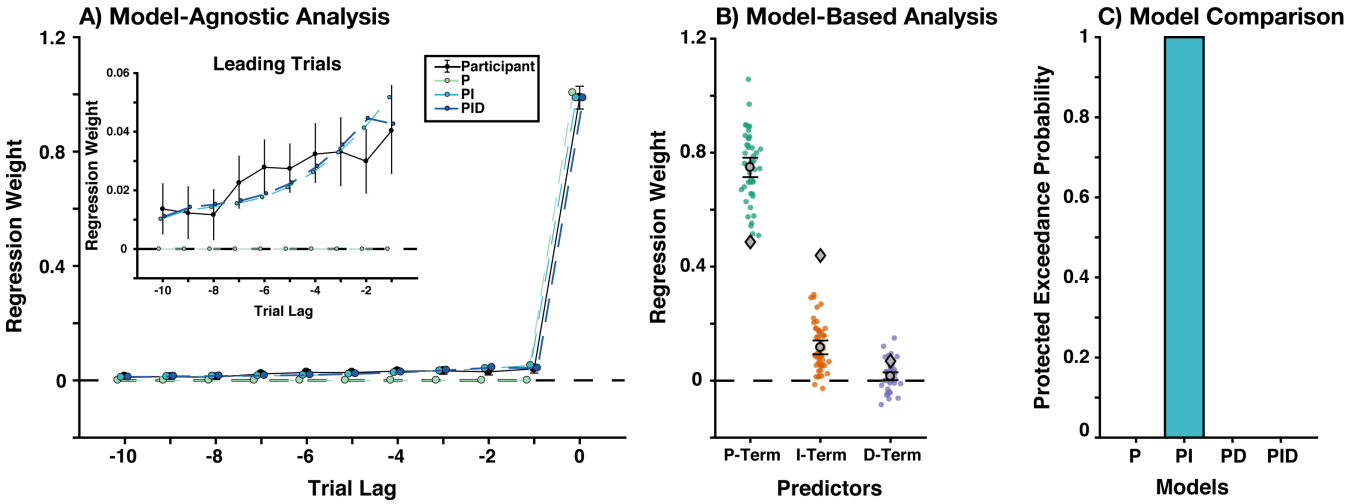

Figure 5. Experiment 3 Results. A) *Participants adjusted their choices based on previous errors, unlike the predictions of a proportional controller (i.e., delta-rule model).* ***B)*** *P, I, and D control significantly predicted participants' updates.* ***C)*** *The PI controller best explained participants' behavior. See Figure 3 for detailed graph legends. Error bars indicate mean and between-participant bootstrapped 95% confidence intervals.*

PID Model Fit

Replicating Experiments 1 and 2, we found that our standard PID model accounted for most of the variance in participants' updates (median $r^2 = 0.81$). The parameters for the P-, I-, and D- terms were all significantly different from zero (standardized beta = Mean(SD); $\beta_P = 0.75(0.12)$, $p \leq 10^{-5}$; $\beta_I = 0.12(0.085)$, $p < 10^{-5}$; $\beta_D = 0.015(0.049)$, $p = .038$; Fig. 5B). The group-level λ was .8020. Participants' estimated gains were similar to the ideal PID controller, but they over-weighted proportional control and under-weighted integral control. We found that there were likely differences between the model likelihoods (Bayesian Omnibus Risk $< 10^{-4}$), and that Bayesian Model Selection strongly favored the PI model ($PXP_{PI} > 0.99$) over the alternate models (all other PXPs < 0.01; Fig. 5C).

Gain Modulation

We examined the independent influence of rewards, absolute error, and outcome entropy in modulating the PID gains across our three experiments. We found that all three modulators significantly interacted with the P, I, and D Terms, but in distinct ways (Fig. 6): increased reward led to an increased P and I gain, and a decreased D gain (Fig. 6A; Mean(SD) interaction beta: $\beta_{P*reward} = 0.082(0.13)$, $p \leq 10^{-5}$; $\beta_{I*reward} = 0.013(0.037)$, $p = .0029$; $\beta_{D*reward} = -0.011(0.033)$, $p = .0052$; sign-randomization test). Increased absolute error led to an increased P gain, and a decreased I and D gain (Fig. 6B; Mean(SD) interaction beta: $\beta_{P*error} = 0.046(0.11)$, $p \leq 10^{-5}$; $\beta_{I*error} = -0.035(0.063)$, $p \leq 10^{-5}$; $\beta_{D*error} = -0.024(0.060)$, $p \leq 10^{-4}$); Increased outcome entropy led to a decreased P gain, and an increased I and D gain (Fig. 6C; Mean(SD) interaction beta: $\beta_{P*entropy} = -0.029(0.063)$, $p \leq 10^{-5}$; $\beta_{I*\,entropy} = 0.064(0.066)$, $p \leq 10^{-5}$; $\beta_{D*\,entropy} = 0.017(0.039)$, $p \leq 10^{-5}$).

These interactions were robust to several quality checks. First, all effects remained significant when corrected for multiple comparisons using the Holm–Bonferroni procedure (Holm, 1979). Given the presence of outliers, we also tested our effects using a robust Wilcoxon signed-rank test (Wilcoxon, 1945), finding that all interactions remained significant (*p*s ≤ .0022). Finally, we found that all interactions remained significant when we did not remove between-experiment variance (*p*s ≤ .0064; Fig. 6 depicts participants' raw interaction betas).

Discussion

In Experiment 3, we found confirmatory evidence that the PID model accurately describes participants' predictions, and that participants adjust their weighting of different PID terms based on trial-wise task dynamics. We found that each PID term

Gain Modulation

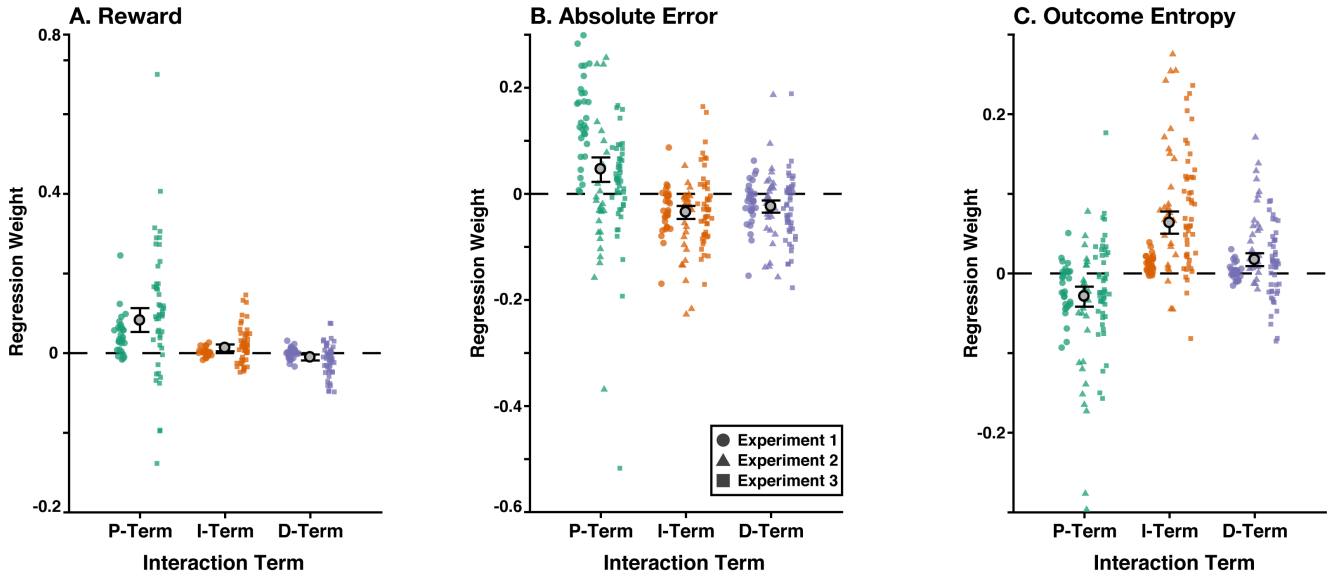

Figure 6. Gain Modulation. *Interaction terms between the trial-wise reward (**A**), absolute error (**B**), and outcome entropy (**C**) with each of the PID terms, for each of the experiments. All models included all main effects. Colored shapes indicate individual participant's standardized betas in each experiment (see legend). Error bars indicate mean and between-participant bootstrapped 95% confidence intervals, uncorrected for between-experiment variance.*

was uniquely sensitive to changes in reward, absolute error, and outcome entropy, extending previous observations of the role of these modulators on proportional control, and providing further evidence that the PID terms represent distinct control processes.

When participants received larger rewards, they modulated their gains in a way that is consistent with a preference for accuracy (P- and I-Terms) over stability (D-term; Ang, Chong, & Li, 2005), and potentially indicating an exploitive strategy for the high reward environments (Kovach et al., 2012). While participants P gain was already larger than the best-performing gain, this may reflect the unique role of reward modulation, controlling for the environmental changes (e.g., entropy) that make a high P gain less desirable. Another alternative is that the P- and/or I-Terms are effortful to implement, with rewards 'paying the cost' of these control policies (Kool, Gershman, et al., 2017; Manohar et al., 2015). Further work will be necessary to dissociate the role of salience and motivation on reward-modulated gain adjustments.

In response to absolute errors (i.e., surprise), participants increased their immediate adjustment (P-Term), and relied less on previous feedback (I- and D-Terms).

This is consistent with the idea that large errors may indicate changes in the environment (Nassar et al., 2010), and with filtering mechanisms in industrial PID control that improve robustness by limiting the long-term influence of noisy samples (Ang et al., 2005).

While outcome entropy, and uncertainty, has traditionally been thought to increase the gain on proportional control (Behrens et al., 2007; Courville et al., 2006; Nassar et al., 2010), in our experiment the P-Term was decreased, and the I- and D-Terms were instead increased. Interestingly, when we instead implement this gain modulation in a P-only model, we do find that outcome entropy increases the gain of the P-term (data not shown). Unlike previous experiments studying uncertainty, environmental change in Experiments 2 and 3 required tracking gradually changing outcomes, which accounted for most of the outcome entropy and for which integral and derivative control are particularly useful (Kovach et al., 2012; Wittmann et al., 2016). In Experiment 1, where these gradual changes were not present, we found that uncertainty after a change-point increased control gains (McGuire et al., 2014), which may be reflected here by integrating over the trials since the change-point.

General Discussion

Across three experiments, we found that the PID model successfully captured participants' prediction updating in a stochastic environment. By incorporating a richer model of control monitoring and adjustment, the PID controller was able to account for ways in which performance in such environments deviates from predictions of standard error-driven (delta-rule) learning models. We also replicated and extended previous findings showing that learning parameters themselves are modulated by environmental signals (e.g., reward), and extended these findings to show that signals related to the magnitude of reward, error, and outcome entropy can differentially affect the gains on the PID model parameters.

Our findings suggest that PID control offers a good account of behavior across two fairly different task environments. Indeed, while we found that normative PID gains differed substantially between Experiment 1 (standard change-point task) and Experiments 2-3 (gradual change-point task), participant behavior continued to qualitatively match the behavior predicted by this normative controller across studies, in each case matching the sign of the best-performing control gain. This suggests that these gains adapted to the specific environment that participants were acting in. Specifically, when outcomes were prone to shift sharply and dramatically (Experiment 1), participants tended to rely less on history-dependent control processes like integral

and derivative control, especially on trials in which large errors may have indicated a state shift.

While we have focused our discussion of the PID controller on all three of its components, in industrial settings the D-Term is often given the lowest gain or not included (Aström & Murray, 2010), as it is highly sensitive to noise. Accordingly, our own data supported little to no role for the derivative term in the current experiments, both normatively and in our model fits to participants' behavior. While the derivative control term was significant in all of the experiments, and interacted with the absolute error, it did not account for sufficient variance to outweigh complexity penalties in model comparison. This may have been compounded the fact that the derivative term was negatively modulated by absolute error, which may have caused it to explain less of the variance on trials where there were large updates. While the outcomes in Experiments 2 and 3 were designed to differentiate PID control from the delta-rule model, they were not designed to specifically detect derivative control. Future research should investigate cases where derivative control is especially beneficial for good performance. Since derivative control provides high-frequency compensation to improve responsivity, it may be the case that derivative control is generally poorly suited for tasks that depend on inter-trial adjustments and favor accuracy over speed. Relative to Experiment 2, Experiment 3 emphasized accuracy through its reward structure, and deemphasized responsivity due to its longer trial length. While there were several differences between these experiments, these factors may have contributed to the differences in derivative control between these experiments.

Some of the most promising evidence for derivative control was that in Experiment 2 participants down-weighted recent errors (from t-3 and t-1) relative to what would be expected by error integration alone. While basic derivative control would only compare the current and previous errors, participant's behavior resembles a common practice in control engineering to low-pass filter the derivative term in order to improve robustness (Ang et al., 2005). The discrepancy between the observed non-linear influence of previous errors (predicted by the full PID model) and the model selection preference for the PI model, may therefore be accounted for by alternative forms of derivative control.

We found that the PID terms were independently modulated by reward (Experiments 1 and 3), absolute error (Experiments 1-3), and outcome entropy (Experiments 1-3) on a trial-to-trial basis. While we have proposed speculative explanations for the role of each modulating factor, at a minimum, the unique pattern of interactions for each of the PID terms suggests that P, I, and D represent dissociable

forms of control. Future experiment should examine the extent to which gain modulation depends on the structure of the task and environment, for instance whether the task rewards consistency in addition to accuracy.

The PID model provides robust control without relying on an explicit model of the environment, offering a parsimonious explanation of the behavior we observed in these experiments. However, there have been notable successes for algorithms that instead learn generative models of the environment (e.g., using Bayesian updating), and can represent the uncertainty about upcoming choices (e.g., Daw, Niv, & Dayan, 2005; Franklin & Frank, 2015; Griffiths, Lieder, & Goodman, 2015; McGuire et al., 2014; Nassar et al., 2010, though see Geana & Niv, 2014; Duverne & Koechlin, 2017). Given the ability of model-based predictions to explain participants' behavior in certain environments, and participants' ability to report confidence in their choices, it is likely that predictions consist of some mixture between model-free control (e.g., PID) and model-based control (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Gläscher, Daw, Dayan, & O'Doherty, 2010; Kool, Cushman, & Gershman, in press; Momennejad et al., 2017). Model-based control may serve to modulate the PID controller itself (e.g., to tune gain parameters or reset control processes; Bouret & Sara, 2005; Behrens et al., 2007; Nassar et al., 2010; McGuire et al., 2014); may be selectively engaged in environments that are highly constrained or familiar; and/or may operate in parallel over longer timescales (Denève, Alemi, & Bourdoukan, 2017). Future experiments should compare the PID controller to model-based approaches, and examine how these models can be combined.

These experiments together provide strong evidence for the viability of control theoretic models as mechanisms of human prediction updating in dynamic environments. This class of models has been highly influential in research on motor control, including the PID controller in particular (e.g., Kawato & Wolpert, 1998). Motor control models typically describe the rapid regulation of limb movements to produce trajectories that are fast, accurate, and robust. In contrast, participants in our experiments were not motivated to make fast or accurate trajectories, and instead may have used analogous control process to adapt their predictions from trial to trial. Control theoretic algorithms (like PID control) may be a domain-general class of neural functions, involved in a diverse array of cognitive processes (Powers, 1973; Pezzulo & Cisek, 2016), including the cognitive control functions that have been suggested to operate using both closed-loop (Botvinick et al., 2001) and open-loop (Shenhav et al., 2013) control principles. The architecture of these executive control algorithms, and the nature of the references that they regulate, are important areas of further research.

Acknowledgments

We would like to thank Kia Sadahiro and William McNelis for their assistance in data collection.

Data Availability

All data and analysis scripts are available upon request.

References

Aben, B., Verguts, T., & Van den Bussche, E. (2017). Beyond trial-by-trial adaptation: A quantification of the time scale of cognitive control. *J Exp Psychol Hum Percept Perform, 43*(3), 509-517. doi:10.1037/xhp0000324

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute, 50*(1), 277-291.

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nat Neurosci, 14*(10), 1338-1344. doi:10.1038/nn.2921

Alexander, W. H., & Brown, J. W. (2015). Hierarchical Error Representation: A Computational Model of Anterior Cingulate and Dorsolateral Prefrontal Cortex. *Neural Comput, 27*(11), 2354-2410. doi:10.1162/NECO_a_00779

Ang, K. H., Chong, G., & Li, Y. (2005). PID control system analysis, design, and technology. *IEEE Transactions on Control Systems Technology, 13*(4), 559-576.

Ashby, W. R. (1956). An introduction to cybernetics. Chapman & Hall.

Aström, K. J., & Murray, R. M. (2010). *Feedback systems: an introduction for scientists and engineers*: Princeton university press.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nat Neurosci, 10*(9), 1214-1221. doi:10.1038/nn1954

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Wiley Series in Probability and Statistics. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, 293-300.

Bernacchia, A., Seo, H., Lee, D., & Wang, X. J. (2011). A reservoir of time constants for memory traces in cortical neurons. *Nat Neurosci, 14*(3), 366-372. doi:10.1038/nn.2752

Blais, C., & Bunge, S. (2010). Behavioral and neural evidence for item-specific performance monitoring. *Journal of Cognitive Neuroscience, 22*(12), 2758-2767.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624.

Bouret, S., & Sara, S. J. (2005). Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. *Trends in neurosciences, 28*(11), 574-582.

Bugg, J. M., & Crump, M. J. C. (2012). In Support of a Distinction between Voluntary and Stimulus-Driven Control: A Review of the Literature on Proportion Congruent Effects. *Frontiers in Psychology, 3*, 367. doi:10.3389/fpsyg.2012.00367

Carter, C. S., Macdonald, A. M., Botvinick, M., Ross, L. L., Stenger, V. A., Noll, D., & Cohen, J. D. (2000). Parsing executive processes: strategic vs. evaluative functions of the anterior cingulate cortex. *Proceedings of the National Academy of Sciences, 97*(4), 1944-1948.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends Cogn Sci, 10*(7), 294-300. doi:10.1016/j.tics.2006.05.004

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences, 10*(7), 294-300.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in quantitative methods for psychology, 1*(1), 42-45.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204-1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci, 8*(12), 1704-1711. doi:10.1038/nn1560

Denève, S., Alemi, A., & Bourdoukan, R. (2017). The brain as an efficient and robust adaptive learner. *Neuron, 94*(5), 969-977.

Duverne, S., & Koechlin, E. (2017). Rewards and Cognitive Control in the Human Prefrontal Cortex. *Cerebral Cortex, 27*(10), 5024-5039.

Franklin, G. F., Powell, J. D., & Emami-Naeini, A. (1994). *Feedback control of dynamic systems* (Vol. 3): Addison-Wesley Reading, MA.

Franklin, N. T., & Frank, M. J. (2015). A cholinergic feedback circuit to regulate striatal population uncertainty and optimize reinforcement learning. *Elife, 4*. doi:10.7554/eLife.12029

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews, 37*(7), 1297-1310.

Geana, A., & Niv, Y. (2014). *Causal model comparison shows that human representation learning is not Bayesian.* Paper presented at the Cold Spring Harbor symposia on quantitative biology.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733-760.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*(4), 585-595.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. *Journal of Experimental Psychology: General, 121*(4), 480.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science, 7*(2), 217-229.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *J Neurosci, 31*(11), 4178-4187. doi:10.1523/JNEUROSCI.4652-10.2011

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2009). Fictive reward signals in the anterior cingulate cortex. *Science, 324*(5929), 948-950.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.

Hunt, L. T., & Hayden, B. Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. *Nature Reviews Neuroscience, 18*(3), 172-182.

Ito, S., Stuphorn, V., Brown, J. W., & Schall, J. D. (2003). Performance Monitoring by the Anterior Cingulate Cortex During Saccade Countermanding. *Science, 302*(5642), 120-122. doi:10.1126/science.1087847

Jiang, J., Beck, J., Heller, K., & Egner, T. (2015). An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands. *Nat Commun, 6*, 8165. doi:10.1038/ncomms9165

Kawato, M., & Wolpert, D. (1998). Internal models for motor control. *Sensory Guidance of Movement, 218*, 291-307.

Kennerley, S. W., Walton, M. E., Behrens, T. E., Buckley, M. J., & Rushworth, M. F. (2006). Optimal decision making and the anterior cingulate cortex. *Nat Neurosci, 9*(7), 940-947. doi:10.1038/nn1724

Kool, W., Cushman, F. A., & Gershman, S. J. (2017). Competition and Cooperation Between Multiple Reinforcement Learning Systems.

Kool, W., Gershman, S. J., & Cushman, F. A. (in press). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, 0956797617708288.

Kovach, C. K., Daw, N. D., Rudrauf, D., Tranel, D., O'Doherty, J. P., & Adolphs, R. (2012). Anterior prefrontal cortex contributes to action selection through tracking of recent reward trends. *J Neurosci, 32*(25), 8434-8442. doi:10.1523/JNEUROSCI.5468-11.2012

Laming, D. R. J. (1968). Information theory of choice-reaction times.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior, 84*(3), 555-579.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & cognition, 7*(3), 166-174.

Manohar, Sanjay G., Chong, Trevor T. J., Apps, Matthew A. J., Batla, A., Stamelou, M., Jarman, Paul R., . . . Husain, M. Reward Pays the Cost of Noise Reduction in Motor and Cognitive Control. *Current Biology, 25*(13), 1707-1716. doi:10.1016/j.cub.2015.05.038

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience, 5*.

Matsumoto, M., Matsumoto, K., Abe, H., & Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. *Nature neuroscience, 10*(5), 647.

McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. *Neuron, 84*(4), 870-881. doi:10.1016/j.neuron.2014.10.013

Mirenowicz, J., & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology, 72*(2), 1024-1027.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *bioRxiv*, 083824.

Nassar, M. R., & Frank, M. J. (2016). Taming the beast: extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences, 11*(Supplement C), 49-54. doi:https://doi.org/10.1016/j.cobeha.2016.04.003

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J Neurosci, 30*(37), 12366-12378. doi:10.1523/JNEUROSCI.0822-10.2010

Niki, H., & Watanabe, M. (1979). Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain research, 171*(2), 213-224.

O'Reilly, J. X., Schuffelgen, U., Cuell, S. F., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proc Natl Acad Sci U S A, 110*(38), E3660-3669. doi:10.1073/pnas.1305373110

Padmala, S., & Pessoa, L. (2011). Reward reduces conflict by enhancing attentional control and biasing visual cortical processing. *Journal of Cognitive Neuroscience, 23*(11), 3419-3432.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*(6), 532.

Pezzulo, G., & Cisek, P. (2016). Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition. *Trends Cogn Sci, 20*(6), 414-424. doi:10.1016/j.tics.2016.03.013

Powers, W. T. (1973). *Behavior: The control of perception*: Aldine Chicago.

Rabbitt, P. (1966). Errors and error correction in choice-response tasks. *Journal of experimental psychology, 71*(2), 264.

Rescorla, R. A., & Wagner, A. W. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99): Appleton-Century-Crofts.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage, 84*, 971-985.

Seo, H., & Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *J Neurosci, 27*(31), 8366-8377. doi:10.1523/JNEUROSCI.2369-07.2007

Shahnazian, D., & Holroyd, C. B. (2017). Distributed representations of action sequences in anterior cingulate cortex: A recurrent neural network approach. *Psychonomic Bulletin & Review*, 1-20.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron, 79*(2), 217-240. doi:10.1016/j.neuron.2013.07.007

Smith, E. H., Banks, G. P., Mikell, C. B., Cash, S. S., Patel, S. R., Eskandar, E. N., & Sheth, S. A. (2015). Frequency-Dependent Representation of Reinforcement-Related Information in the Human Medial and Lateral Prefrontal Cortex. *J Neurosci, 35*(48), 15827-15836. doi:10.1523/JNEUROSCI.1864-15.2015

Tervo, D. G., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., & Karpova, A. Y. (2014). Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell, 159*(1), 21-32. doi:10.1016/j.cell.2014.08.037

Ullsperger, M., Danielmeier, C., & Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiol Rev, 94*(1), 35-79. doi:10.1152/physrev.00041.2012

Wang, X.-J. (2008). Decision making in recurrent neuronal circuits. *Neuron, 60*(2), 215-234.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits*. Retrieved from IRE WESCON Convention Record:

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin, 1*(6), 80-83.

Wilkinson, G., & Rogers, C. (1973). Symbolic Description of Factorial Models for Analysis of Variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 22*(3), 392-399. doi:10.2307/2346786

Wittmann, M. K., Kolling, N., Akaishi, R., Chau, B. K., Brown, J. W., Nelissen, N., & Rushworth, M. F. (2016). Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex. *Nat Commun, 7*, 12327. doi:10.1038/ncomms12327

Yu, A. J., & Dayan, P. (2005). Uncertainty, Neuromodulation, and Attention. *Neuron, 46*(4), 681-692. doi:https://doi.org/10.1016/j.neuron.2005.04.026